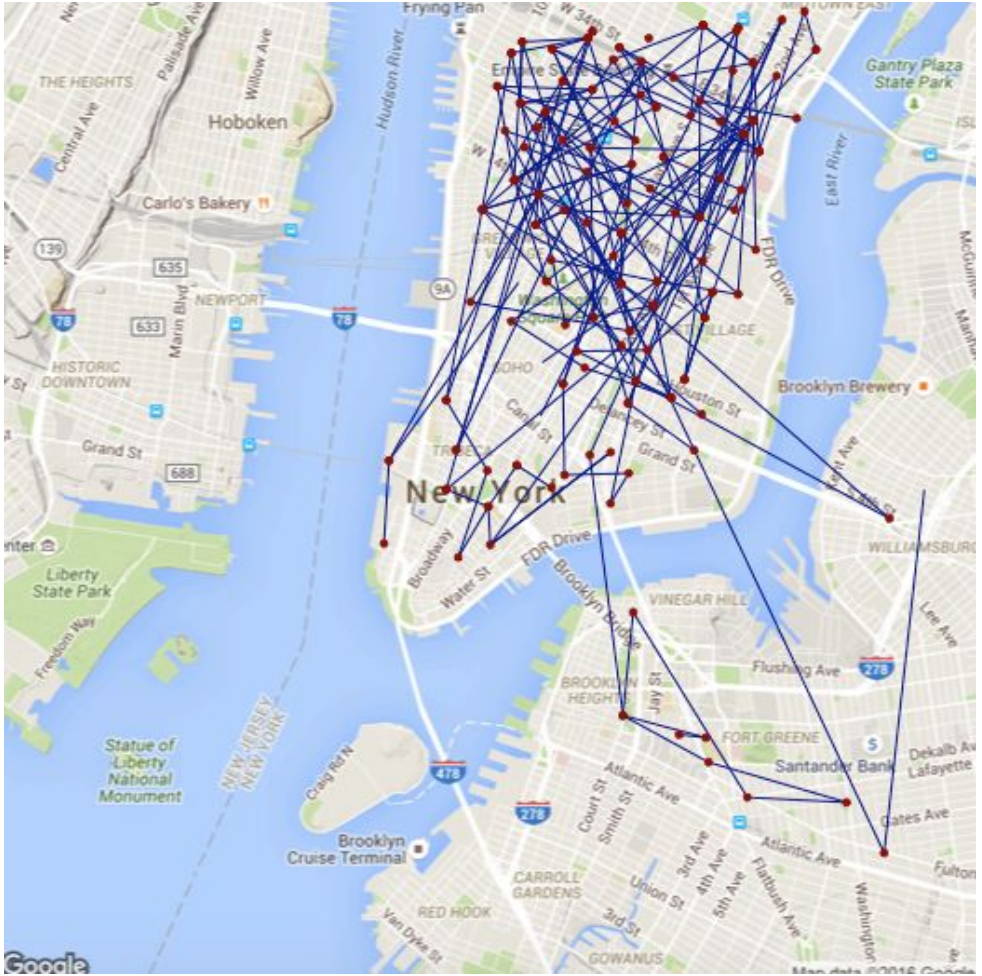
Analyzing NYC's Citi Bike Data

Jonathan Landesman and Cory Kind
April 2016

About the Data

NYC's Citi Bike program debuted in 2013. Since then, over 24M bike rides have been completed between 491 stations (and counting!) in New York.

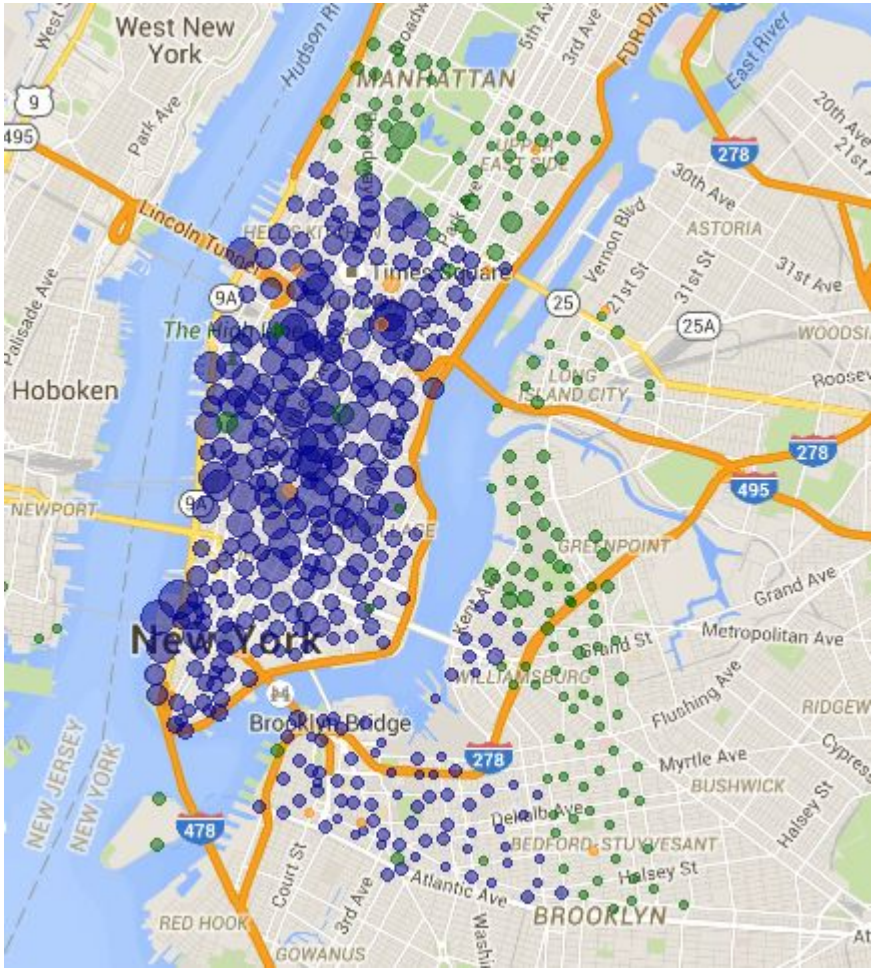
The full data is available through NYC Open Data.



About the Data

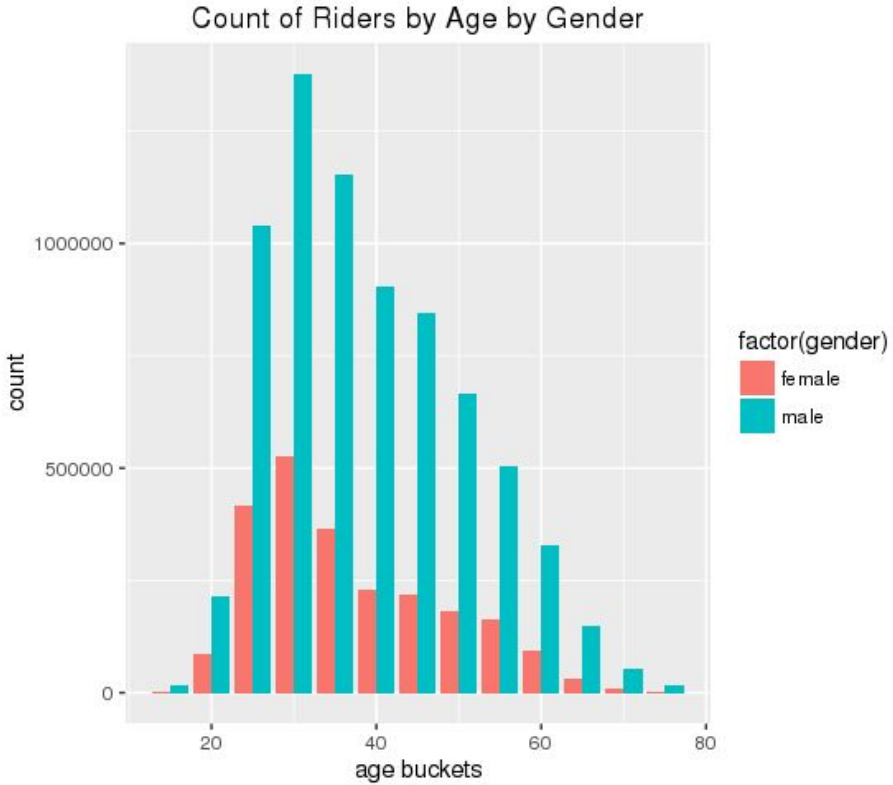
This map shows the growth of the Citi Bike program over time. Larger bubbles are more active stations.

- Blue: Introduced in 2013
- Green: Added as of 2015
- Orange: Added as of Feb. 2016

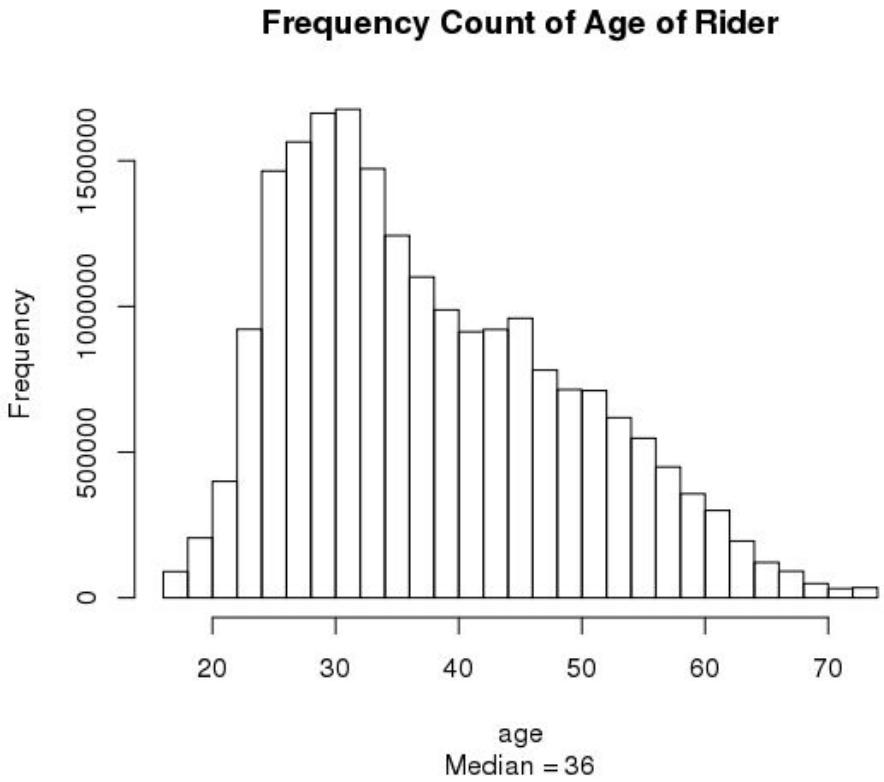


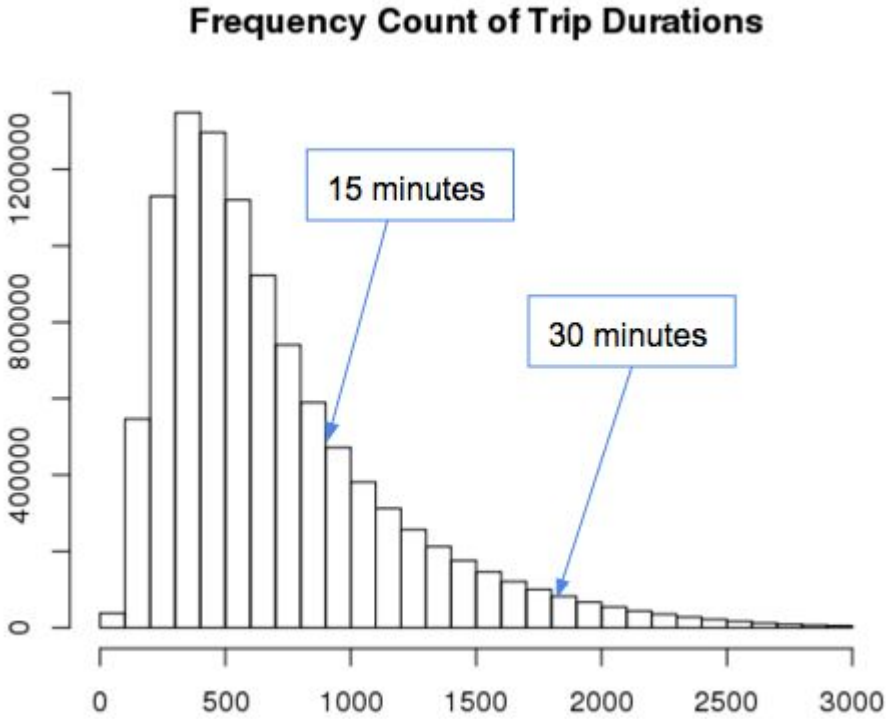
Data Overview: Age and Gender

- Trips by male riders far outnumber trips by female riders
- May indicate gender imbalance in ridership, imbalance of number of rides/rider, or both.
- 25-35 are peak ages for both genders



Data Overview





Primary Question

How does age affect cardiovascular ability among users of New York City's Citi Bike program, as demonstrated by average biking speed between stations?

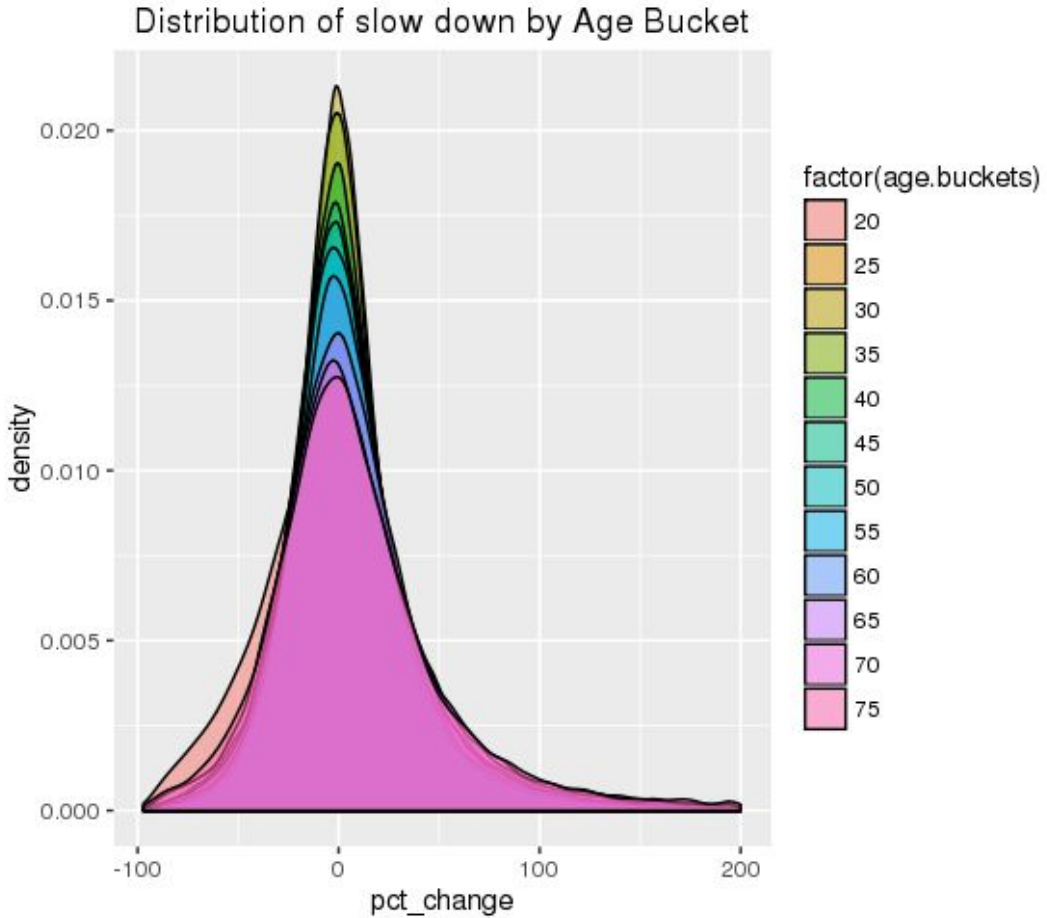
This information could be used to understand the effects of regular biking on cardiovascular health and provide support for bike share programs.

Analysis Plan

- Clean and filter data, assign records to 5-year age buckets
- Group by path (start station -> end station), gender, month, and age bucket, and find the average trip duration for each group
- For each group, calculate the difference between that group's average trip duration and the baseline for that path and gender.
- For each group, calculate the differential (the percentage difference between the average trip duration and the baseline for that path and gender).
- Aggregate up the differentials.

Results (Preliminary)

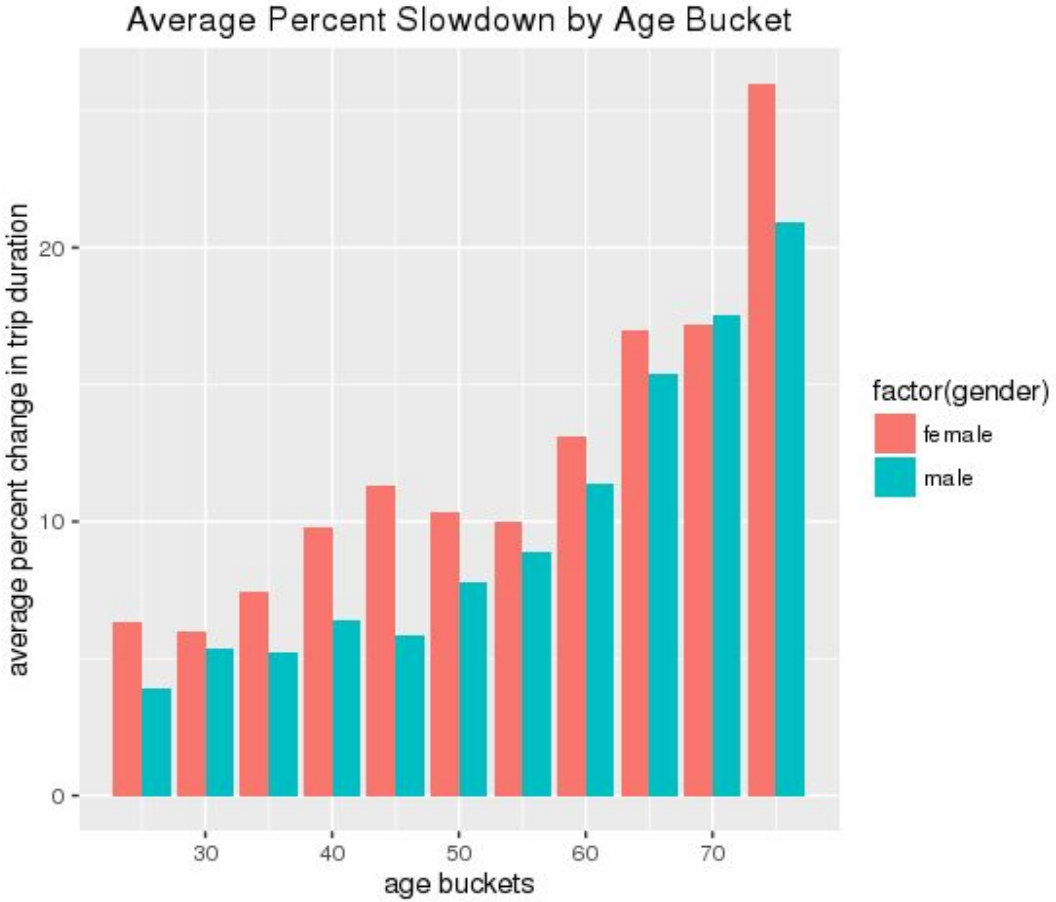
Differential approach supported earlier linear model analysis - age has a dramatic impact on biking speed, with older riders spending longer on the same routes.



Results (Preliminary)

The percentage slowdown increases with age, as expected.

The slowdown appears to be larger for women than men, but final data will present a clearer picture.



Caveats

- **No causality:** Without an experimental design, unidentified factors could be leading to the results we observed. For instance, older people may be more likely to stop on the way or take a scenic route to avoid traffic.
- **Selection bias:** Citi Bike subscribers are self-selecting. They will probably skew healthier than the general population. This bias will likely increase with age, as more and more people with health issues opt out of Citi Bike membership.

Architecture

- **Infrastructure**
  - AMI: 205UCB_RStudio
  - M2.4 X-large instance (68.4 GiB memory)

- **Software**
  - Postgres 8.4.20 for storage
  - R for analyzing data, via RPostgreSQL (dplyr, ggplot2, lubridate, ggmaps for analysis)

Lessons Learned: Analysis Plan

- **Original plan:** Calculate differentials among people who took the same path within the same day or even hour, to account for weather and traffic.
- **Why it didn't work:** Not enough cases. 2.4M rides spread over 3 years and 100K possible paths from Station A -> Station B means that we don't have enough riders per hour or day to get that level of granularity.
- **What we did instead:** We assumed that weather and traffic balanced out over time, and disregarded day and time in the analysis. The differentials are grouped by month.

Lessons Learned: Infrastructure

- **Original plan:** Load data into HDFS, load into Hive, and then run scripts in R using the Hive package.
- **Why it didn't work:** R is in-memory and does not automatically cache. Hive queries were taking much longer than we had assumed.
- **What we did instead:** Because our schema was fixed ahead of time, we used Postgres as our storage solution rather than Hive and used the RPostgreSQL to load data into R. That sped up the queries dramatically.

Future Questions - Where to go from here?

- Incorporate other open datasets to account for traffic and weather patterns.
- Incorporate data from real-time JSON feed.
- Additional Analysis Questions:
  - How has the use of the Citi Bike program has changed over time? Have the demographics shifted as the program has expanded?
  - What can we learn about the city from Citi Bike data? At which stations is Citi Bike usage expanding quickly? Are bikeshare stations a precursor of neighborhood economic growth, as some research has argued?